**Selection with BLUP**

Jeffrey Endelman

Agron/Hort 812, Spring 2020

From a set of *n* selection candidates, we wish to select the *k* < *n* genotypes with the highest genotypic values for a trait (or index of multiple traits).

Assume the genotypic values are normally distributed, with known mean and variance,

but they are not directly observable. The genotypic values must be predicted from measured phenotypes. The ideal prediction method would give the same ranking of genotypes as the true values.

Assume a Completely Randomized Design, but the number of replicates (*mi* for genotype *i*)is not the same for all genotypes, either intentionally (partially replicated design) or because some are missing at random.

|  |  |
| --- | --- |
| = phenotype for replicate *j* of genotype *i* | Eq. 1 |

**How to predict genotypes?**

Most intuitive answer is to use the average phenotype for each genotype:

This is the ordinary least squares (OLS) estimator, i.e., it minimizes the sum of squared residuals (SSR):

However, OLS does not provide an optimal ranking of genotypes when the dataset is unbalanced.

The problem of finding the optimal ranking when different numbers of observations are available for each genotype was one of the primary motivations for Henderson, who was an applied statistician working in the field of animal breeding, to develop BLUP in the 1950's and 60's. My initial treatment of this topic is technically called BLP, not BLUP. Both assume the variances are known, but the former also assumes the mean is known and is therefore simpler to present.

The BLP solution to our simple selection experiment is

|  |  |
| --- | --- |
|  | Eq. 2 |

where is the broad-sense heritability on an entry-mean basis:

|  |  |
| --- | --- |
|  | Eq. 3 |

BLP shrinks the phenotypic average toward the mean, and the amount of shrinkage is given by the heritability, which is smaller for genotypes with fewer replicates.

**Properties of BLP**

The following are general properties of BLP (not just for the example in Eq. 1):

1. It maximizes the probability of correctly ranking two genotypes:

1. It maximizes the expected genotypic value of the selected group *Sk*:
2. It minimizes the variance of the prediction error:
3. It maximizes the correlation between the true and predicted genotypic values:

|  |  |
| --- | --- |
|  | Eq. 4 |

The second equality in Eq. 4 follows from the property:

which differs from the OLS result:

The squared correlation between the true and predicted genotypic values is called **reliability** in the animal breeding literature. By substituting Eq. 2 into Eq. 4, one finds that reliability is identical to broad-sense heritability for the completely randomized design:

This equivalence also holds for multi-environment trials.

**Matrix formulation**

To analyze more complex experimental designs, or when genomic relationships are modeled, a multivariate formulation of BLUP is helpful. To gain experience with matrix notation, consider a simple example of the above CRD problem involving two genotypes, with one replicate of genotype 1 and two replicates of genotype 2. The matrix equivalent of Eq. 1 is

|  |  |
| --- | --- |
|  | Eq. 5 |

According to Eqs. 2 and 3, the best linear predictor in this example is

|  |  |
| --- | --- |
|  | Eq. 6 |

A general formulation of the random effects model is

|  |  |
| --- | --- |
|  | Eq. 7 |

The boldface font indicates either a vector (lowercase) or matrix (uppercase). The variance-covariance matrix for the phenotypes is

|  |  |
| --- | --- |
|  | Eq. 8 |

The general formula for the best linear predictor is

|  |  |
| --- | --- |
|  | Eq. 9 |

For the example in Eq. 5,

|  |  |
| --- | --- |
|  | Eq. 10 |

In the homework, you are asked to derive Eq. 6 from the general formula in Eq. 9.

**Mixed Models and BLUP**

Mixed models have both fixed (and random (**u**) effects, and it is conventional (but not necessary) to assume the random effects have zero mean:

|  |  |
| --- | --- |
|  | Eq. 11 |

If were known, the best linear predictor of **u** would follow from Eq. 9:

|  |  |
| --- | --- |
|  | Eq. 12 |

However, if the fixed effects are unknown, the best predictor (called BLUP) is calculated by replacing with its Generalized Least Squares (GLS) estimate:

|  |  |
| --- | --- |
|  | Eq. 13 |

One disadvantage to computing BLUP according to Eq. 13 is that it requires inversion of **V**, which depends on the number of phenotypes. A more computationally efficient procedure (due to Henderson) is to solve the following linear system, called the mixed model equations (MME):

|  |  |
| --- | --- |
|  | Eq. 14 |

Although **R** has the same dimensions as **V**, it is typically diagonal and therefore easily inverted. The computational complexity of solving Eq. 14 depends on the number of genotypes rather than the number of phenotypes.

Reliability was introduced in Eq. 4 as the squared correlation between the true and predicted values. For the more general model in Eq. 11, reliability can be calculated from the inverse of the coefficient matrix in the MME (Eq. 14), which has a block-diagonal structure:

|  |  |
| --- | --- |
|  | Eq. 15 |
|  | Eq. 16 |

In Eq. 16, is the ith diagonal element of the **C**22 matrix in Eq. 15, and **D***ii* is the ith diagonal element of **D**.

**References (not required reading)**

Fernando RL and Gianola D. 1986. Optimal properties of the conditional mean as a selection criterion. Theor Appl Genet 72:822–825.

Henderson CR. 1963. Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*. (Hanson WD and Robinson HF, eds.) pp. 141-63. National Academy of Sciences and National Research Council Publication No. 982, Washington, DC.

Searle SR. 1973. Prediction, mixed models, and variance components. Biometrics Unit Technical Report BU-468-M.